FREE RESOURCE FROM FUTURE FACTORS AI

The AI Trust Checklist

Five checks to run on an AI assistant before you scale it, and a way to score the result. By Hina Mian, Co-Founder of Future Factors AI.

How to use this checklist

What it is. A set of five checks, each with questions to ask, evidence to look for, a red flag and an owner. Together they help you decide whether an AI assistant, or a single answer from one, is safe to rely on and safe to scale.

Who fills it in. A team before launching an assistant, with each check completed by the owner named for it. An individual can use the 60-second check in the article before relying on one output.

How to answer. Tick Yes, No or Unsure for each of the 20 items on pages 2 and 3, name the owner, and write down the evidence you looked at. Unsure is not a polite No. It means you do not have evidence yet.

How to score. Yes = 2, Unsure = 1, No = 0. The maximum is 40. Four items marked with a star (1.1, 2.3, 3.3, 5.2) are must-haves.

Band	Score	Must-haves	What to do
GREEN	34 to 40	All answered Yes	Go to a wider pilot.
AMBER	24 to 33	Or any must-have answered Unsure	Small pilot only, a human reviews every output, fix the gaps first.
RED	Under 24	Or any must-have answered No	Do not scale. Fix first.

What this is, and is not. The AI Trust Checklist is a Future Factors working tool, built from published frameworks and practical experience. It has not been validated or tested as a measurement instrument. Scores are a structured way to start a conversation and make a go or no-go call. They are not a certification, a rating or proof of compliance, and this is not legal advice. The thresholds are our judgment, and your team should adjust them. Rules vary by region and by use, so ask your privacy lead or qualified counsel what applies to you.



Pages in this pack

Page 1: how to use it
Pages 2 to 3: the 20-item checklist
Page 4: trust label to fill in
Page 5: pilot one-pager

Read the full guide at futurefactors.ai/ai-trust-checklist/

The checklist: 20 items, five checks

Tick Yes, No or Unsure. A star marks a must-have. Write the evidence you actually looked at.

1 PRIVACY AND CONSENT BOUNDARIES

Owner: Data owner or privacy lead, with IT for system access
Ask these questions:
- What data about people does the assistant read, and from which systems (CRM, LMS, HR system, shared drives)?
- Why does it need each field, and does the purpose need it at all?
- Do the people affected know, in plain words, and can they ask, correct or opt out?
- Which regional rules apply to the people whose data this touches?

Red flag: 'It uses everything in the CRM' or users first learn what it knows about them when it quotes it back.

Item	Yes	No	?	Owner (role)	Evidence (what we looked at)	Notes
*** MUST-HAVE** **1.1** We can list every system the AI can read.	[ ]	[ ]	[ ]			
1.2 For each source we know which fields about people it uses and why.	[ ]	[ ]	[ ]			
1.3 We have confirmed in writing whether the vendor stores our prompts or trains on our data.	[ ]	[ ]	[ ]			
1.4 People are told in plain words what is used about them, and how to question it.	[ ]	[ ]	[ ]			

2 ACCURACY AND ATTRIBUTION

Owner: Knowledge owner (L&D lead, content owner or subject expert)
Ask these questions:
- Which sources is the assistant allowed to use, and who approved them?
- Is outdated or unvetted material (an old LMS course, a superseded policy) kept out or clearly labelled?
- Can a person see where each answer came from and click through?
- Who re-tests it after a source or model change?

Red flag: Citations nobody can open, or a 2019 course sitting beside this year's policy with nothing to tell them apart.

Item	Yes	No	?	Owner (role)	Evidence (what we looked at)	Notes
2.1 We have an approved source list; unvetted or outdated material is excluded or labelled.	[ ]	[ ]	[ ]			
2.2 Every source has a named owner and a last-reviewed date.	[ ]	[ ]	[ ]			
*** MUST-HAVE** **2.3** Answers show their sources so a person can check them.	[ ]	[ ]	[ ]			
2.4 We test with real questions that have known answers before launch and after any change.	[ ]	[ ]	[ ]			

3 FAIRNESS AND HARM PREVENTION

Owner: People or HR lead, with the use-case owner
Ask these questions:
- Who could this disadvantage (roles, regions, languages, accessibility needs, people with less data on file)?
- Do we compare outputs for those groups using the same prompt?
- Does it make or quietly drive decisions about hiring, pay, promotion, discipline or assessment?
- What must it never say or do, and how does someone report a problem?

Red flag: A score or ranking about people goes straight into a decision with no human choice and no way to contest it.

Item	Yes	No	?	Owner (role)	Evidence (what we looked at)	Notes
3.1 We have named who could be disadvantaged by this use.	[ ]	[ ]	[ ]			
3.2 We compare outputs across those groups before launch and on a schedule.	[ ]	[ ]	[ ]			
*** MUST-HAVE** **3.3** A person makes any decision about someone's job, pay or assessment, and that person can be challenged.	[ ]	[ ]	[ ]			
3.4 We have a written list of things the assistant must not do, and a way to report problems.	[ ]	[ ]	[ ]			

4 BRAND AND CULTURAL CONSISTENCY

Owner: Brand or communications lead, with the L&D design lead
Ask these questions:
- How should each assistant sound, and who decided (a sales coach can be formal, a leadership coach conversational, on purpose)?
- Do all our tools give the same answer to the same policy question?
- Are the language and examples right for the people using it, in their region and culture?
- Who samples outputs against the voice guide?

Red flag: The same question gets a different policy answer in two tools, or the assistant's tone changes depending on who asks.

Item	Yes	No	?	Owner (role)	Evidence (what we looked at)	Notes
4.1 Each assistant has a written voice guide that someone chose on purpose.	[ ]	[ ]	[ ]			
4.2 The same core questions get consistent answers across our tools.	[ ]	[ ]	[ ]			
4.3 Examples, idioms and spelling suit the audience and region.	[ ]	[ ]	[ ]			
4.4 Someone samples outputs against the voice guide on a schedule.	[ ]	[ ]	[ ]			

5 HUMAN-IN-THE-LOOP AND AUDITABILITY

Owner: Use-case owner or manager, with IT or security for logs
Ask these questions:
- For each type of output, does a person review before use, after use or by sample, matched to what happens if it is wrong?
- Can a user question an answer and escalate to a named person?
- Is there a log of who asked what, which sources were used and when?
- Can a user see why an answer was given, and can we pause the assistant?

Red flag: Nobody can say who reviews outputs, or a wrong answer cannot be traced back to its sources.

Item	Yes	No	?	Owner (role)	Evidence (what we looked at)	Notes
5.1 We matched human review to the consequence of an error for each output type.	[ ]	[ ]	[ ]			
*** MUST-HAVE** **5.2** Users can question an answer and escalate to a named person.	[ ]	[ ]	[ ]			
5.3 We keep a log of use and sources, and we know how long it is kept.	[ ]	[ ]	[ ]			
5.4 Users can see why an answer was given, and we can pause the assistant.	[ ]	[ ]	[ ]			

Total: ____ / 40 **Must-haves all Yes?** [ ] Yes [ ] No **Band:** [ ] Green [ ] Amber [ ] Red

Trust label: what this assistant uses, what it does not, and who to ask

Fill this in for one assistant and show it wherever people use it. Write for the person using the assistant, in plain words, specific enough that someone could prove it wrong.

What it is for	
What it uses (data and sources, with last-reviewed date)	
What it does not use or do	
How it was checked	
How it sounds	
Where a human steps in	
Who to ask (named role and how)	
Last updated	

If you operate in the EU, check whether Article 50 of the AI Act sets transparency duties that cover your assistant. This is not legal advice.

Pilot one-pager

One use case, one owner. Complete this before the pilot starts, and again at each review.

Use case	
Owner (role)	
Data used	
Approved sources	
Rules	
Tone	
Human steps	
Escalation contact	
Success signal	
Review date	
Go or no-go	

Judge the pilot in order: **Use** (are people asking it real questions?), **Persistence** (a month later, are they still using it and still opening the sources?) and **Impact** (is the work better, faster or different?).